High variation subarctic topsoil pollutant concentration prediction using neural network residual kriging

Cite as: AIP Conference Proceedings **1836**, 020023 (2017); https://doi.org/10.1063/1.4981963
Published Online: 05 June 2017

A. P. Sergeev, D. A. Tarasov, A. G. Buevich, I. E. Subbotina, A. V. Shichkin, M. V. Sergeeva, and O. A. Lvova

High Variation Subarctic Topsoil Pollutant Concentration Prediction Using Neural Network Residual Kriging

Sergeev A.P. [1, 2, a)], Tarasov D.A. [1, 2, b)], Buevich A.G. [1, c)], Subbotina I.E [1, d)], Shichkin A.V.[1, 2, e)], Sergeeva M.V.[1, f)], and Lvova O.A. [2, g)]

[1] *Institute of Industrial Ecology UB RAS, Kovalevskoy, 20, Ekaterinburg, RUSSIA 620990.*
[2] *Institute of Radio-electronics and IT, Ural Federal University, Mira, 19, Ekaterinburg, RUSSIA 620002.*
[a)] *Corresponding author: alexanderpsergeev@gmail.com*
[b)] *datarasov@yandex.ru*
[c)] *bagalex3@gmail.com*
[d)] *iesub@mail.ru*
[e)] *and@ecko.uran.ru*
[f)] *marin@ecko.uran.ru*
[g)] *olvova@bk.ru*

Abstract. The work deals with the application of neural networks residual kriging (NNRK) to the spatial prediction of the abnormally distributed soil pollutant (Cr). It is known that combination of geostatistical interpolation approaches (kriging) and neural networks leads to significantly better prediction accuracy and productivity. Generalized regression neural networks and multilayer perceptrons are classes of neural networks widely used for the continuous function mapping. Each network has its own pros and cons; however both demonstrated fast training and good mapping possibilities. In the work, we examined and compared two combined techniques: generalized regression neural network residual kriging (GRNNRK) and multilayer perceptron residual kriging (MLPRK). The case study is based on the real data sets on surface contamination by chromium at a particular location of the subarctic Novy Urengoy, Russia, obtained during the previously conducted screening. The proposed models have been built, implemented and validated using ArcGIS and MATLAB environments. The networks structures have been chosen during a computer simulation based on the minimization of the RMSE. MLRPK showed the best predictive accuracy comparing to the geostatistical approach (kriging) and even to GRNNRK.

Keywords: Artificial Neural Networks, Chromium, Pollution, Residual kriging, GRNNRK, MLPRK

1. INTRODUCTION

Predicting the distribution of soil pollutants is a substantial area of research given the current concerns regarding environmental issues worldwide. Because of the risk to health and environment associated with gain in soil pollution, it is essential to have a model that is able to precisely predict the distribution of pollutants within analyzed territory. Moreover, the problem of prediction the distribution of the element with high variability in the concentration at the study site is particularly difficult. Rapid industrialization over the last decades has significantly contributed to the gain in soil contaminants in Arctic and sub-Arctic regions of Russia. Thus, studies on the impact of urban environment on the soil pollution in the Arctic and sub-Arctic keep on being important fields of investigation.

Modelling might be the method that facilitate the location and delineate the pollution origin sources. Interpolation is one of the most widely used modeling methods. Geostatistical interpolation techniques (e.g. kriging) utilize the statistical features of the measured spots together with the spatial autocorrelation between them and account for the spatial configuration of the sample spots at the prediction location. Kriging has shown considerable

Applied Mathematics and Computer Science
AIP Conf. Proc. 1836, 020023-1–020023-6; doi: 10.1063/1.4981963
Published by AIP Publishing. 978-0-7354-1506-5/$30.00

advantages in the prediction of soil properties, compared with deterministic methods (Schloeder et al., 2001; Liu et al., 2008; Worsham et al., 2010), however high pollution heterogeneity requires more efficient methods.

Nowadays, the famous modeling technique is artificial neural network (ANN). A brief overview of ANN (Bishop, 1995) showed how ANN can be generally applicable. Artificial intelligence methodologies can help to forecast the pollutants in complicated non-linear contexts. The ANN model might be applied to the measured data obtained in the monitoring, and can be used to predict the pollutants content at unmonitored locations (Kanevski, 1999; Liu et al., 2009). Moreover, a combination of different methods is capable to neutralize their weaknesses and to multiply their dignities. In particular, integrating ordinary kriging of residuals from ANN can incorporate the spatial autocorrelation of measured values which can lead to better predictions and lower error.

In this work, we propose two hybrid models combining the techniques of ANN based forecasting and kriging. We examine the results obtained by applying the models to predict the levels of the pollutant (Cr) at a particular location in the sub-Arctic Novy Urengoy, Russia using previously obtained data on pollutant's content as inputs. We also compare the models application with those from the previously conducted neural networks (GRNN and MLP) content prediction.

2. METHODS

Both methods applied (GRNNRK and MLPRK) are three-step algorithms combining two interpolation techniques of ANN and kriging. The first step implies estimating large-scale nonlinear trends using neural networks. The second step is analysis of the stationary residuals by ordinary kriging, which is able to provide local estimates. The final step is estimation produced as a sum of ANN predictions and ordinary kriging estimates of the residuals. In the work, the ANNs were carried out in MATLAB using the GUI interface; the ArcGIS application was performed to predict the values by kriging. The input data set (150 data points) was randomly divided on two subsets: training data set (105 samples) and test data set (45 samples). Training data set was used for building kriging and training ANNs. The predictive accuracy of each selected approach was verified by the correlation coefficient, MAE (1) and RMSE (2) between the prediction and initial data from the training data set.

$$MAE = \frac{\sum_{i=1}^{n} |y_{Modi} - y_i|}{n} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_{iMod} - x_i)^2}{n}} \quad (2)$$

GRNN Approach

GRNNs are known as variation of the radial basis functions (RBF) neural networks, for which a hidden layer is centered at every training sample, and a universal approximator for smooth functions. The structure of basic GRNN is shown in Figure 1.a. The first layer in GRNN resembles the RBF with the amount of neurons that equivalent to the quantity of input vectors. Training of a GRNN is performed in one pass of the training data through the network. Therefore it is fast. Generally, the network has four layers of neurons: input, pattern, summation and output. GRNN, like RBF network has a radially base layer with the number of neurons, equal to or less than the number of elements of the training data set, but also includes a linear layer. The network copies inward all training observations and uses them for estimation the response in an arbitrary point. Final output network evaluation is obtained as a weighted average of outputs over all training observations, where the weights values represent the distance between these observations and the point in which the evaluation is made. Thus, closer points more contribute to the estimation. The first layer of a GRNN consists of radial elements. The second layer (linear) contains elements that help to evaluate the weighted average. This is done using a special procedure. Each output in the layer has its own element forming a weighted sum for it. In order to obtain the weighted average from a weighted sum, this sum should be divided into a sum of weight coefficients. The latter is calculated by a special element of the second layer. After that, the division is produced in the output layer by special "division" elements. Thus, the number of elements in the second layer is one greater than in the output layer. Typically, approximation issues require estimating the only one output value and consequently, the second layer comprises two elements. The learning process of GRNN is similar to RBF one. Initially, the basis functions centers are configured, and then the output layer learns with fixed parameters of RBF neurons. The neurons in the pattern layer perform a nonlinear transformation of the input vectors. Choosing the *spread* parameter of the RBF, which is known as a smoothing parameter, determines the width of the input area, to which each basis function responds. It is the distance from the center of a Gaussian where

the value is one-half of the peak value. In our case, the network has 105 input neurons according 105 sampling points formed the training data set. During the simulation, the spread parameter varied from 0 to 0.3 with step 0.01; in total, 300 simulations were done. The minimal RMSE was achieved with spread parameter of 0.031.

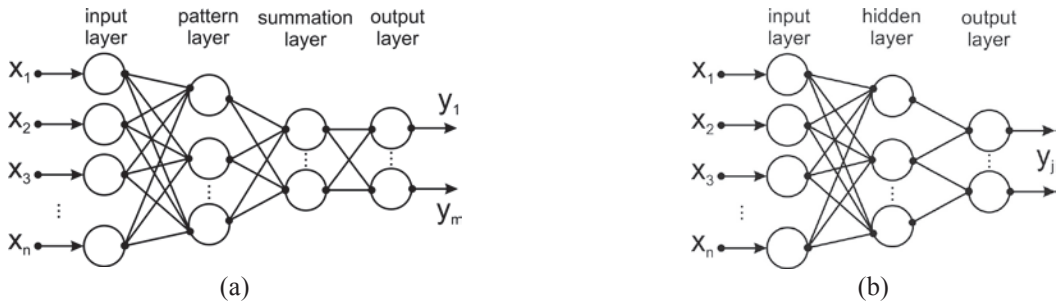

FIGURE 1. ANN structures: a) GRNN; b) MLP

MLP Approach

The most frequently used ANN in environmental studies is MLP (see Figure 1.b), a feed forward network with back propagation. Due to the wide distribution, this type of network is well developed and has shown its high performance. The network structure is described by several numbers relating to the number of neurons in layers: input layer – hidden layer – output layer (for example, 5–3–1 for three-layer-perceptron, 5 input neurons, 3 hidden neurons in one layer, 1 output neuron). If there is more than one hidden layer, the central unit might consist of several numbers. The activation of the neurons is given by the following expression: $a_j = \varphi(\sum_{i=0}^{d} w_{ji}\, x_i$. The outputs of the ANN y are described as $y_k = \varphi(\sum_{j=0}^{m} w_{kj}\, a_j$ where w_{kj} denotes a weight in the second layer connecting hidden neuron j to output neuron k. Combining the expressions above, the complete expression for the transformation of the network is determined as $y_k = \varphi(\sum_{j=0}^{m} w_{kj}\, a\, (\sum_{j=0}^{d} w_{ji}\, x_i)$. In order to train the network, its mapping function y must be differentiable. The amount of activation functions is large and depends on the type of network, number of neurons, solving problem, etc. For the reason, a sigmoidal activation function φ is often used. In practice, a convenient choice is the "tanh" function.

The network structure was determined during computer simulation. In our case, the input layer of MLP was compiled with sampling points; the hidden layer consisted of a few neurons, and the output layer representing the element content in the relevant sample. The selection of the number of neurons in the hidden layer was carried out by the lower total RMSE of prediction of the pollutant (Cr) content for the training (105 samples), test (45 samples), and a complete set of data (150 samples). The number of neurons was varied from 2 to 25. Each network was trained by 500 times and the best of them have been selected. The final configuration of the network selected was 1-5-1, e.g. the hidden layer contains 5 neurons. Network education quality was checked by the correlation coefficient and RMSE between the results of the network predictions and the training data set.

Residuals Implementation

The starting procedure for the residual kriging is the prediction of residuals by the neural network in the test points. Residuals in the neural network can be defined as follows: $r(x_i) = Z(x_i) - Z_{ANN}(x_i)$, where $r(x_i)$ – the residuals of data set x_i, $Z(x_i)$ - measured values, $ZANN(x_i)$ - value estimated by the neural network. The resulting residuals were estimated using kriging. Evaluation in ordinary kriging (OK) is constructed as a linear combination of input data: $r_{OK}(x) = \sum \lambda_i\, r(x_i)$, where $r_{OK}(x)$ – the estimated value at the point x using OK, $\lambda_i(x)$ – the optimal weights with the condition $\Sigma\lambda_i = 1$, and $r(x_i)$ – the residual of a neural network for the point x_i. The OK in ArcGIS application was used in order to predict the research field's residuals. The final evaluation of the pollutant content $Y(x_i)$ was obtained as the sum of the neural network evaluation and residual evaluation by kriging. $Y(x_i) = Z_{ANN}(x_i) + r_{OK}(x_i)$. So that verify the method proposed in the study, a comparison with a stochastic interpolation method Universal Kriging was carried out, then the accuracy of predictions were compared.

3. STUDY CASE

Data for the study were obtained from the results of the soil survey in Novy Urengoy (N66.084722°, E76.678889°), Yamalo-Nenets Autonomous Okrug, Russia (Sergeev et al., 2015), where a chromium anomaly was described. The area of sampling was approximately 8.5 km^2 (see Figure 2). The terrain was flat and covered with peaty-podzolic-gley illuvial-humus sand soil. In total, 150 topsoil samples were collected. The detailed spatial location of sampling points is shown in Figure 2 (b). All the samples were randomly split into independent training and validation (test) data sets. The training data set (105 samples) was used for training the GRNN and interpolating surface pollutant distribution, and the validation (test) data set (45 samples) was used only as independent data set for testing. Concentration indicators for the element (Cr) were obtained by chemical analysis. The descriptive statistics of modeled elements are shown in Table 1.

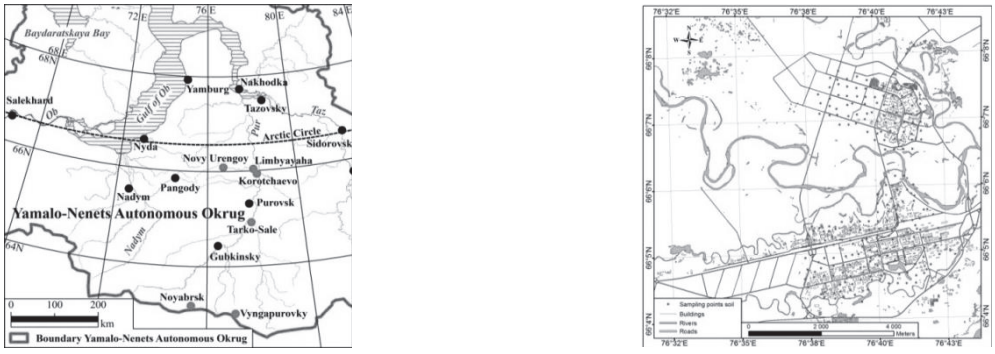

FIGURE 2. The sampling place: a) Yamalo-Nenets Autonomous Okrug, Russia; b) Novy Urengoi city

TABLE 1. Descriptive statistics of the modeled element (*Cr*)

Pollutant	Min	Max	Mean	SD	CV	Skewness	Kurtosis	Median
Cr	25.8	1265.4	245.2	256.3	382.9	1.41	4.61	89.5

From the basic statistics table, it is observed that the element attributes are erratic and positively skewed in nature. Mean concentration of total Cr at the anomaly spots was about ten times higher than at the urban background. Comparing to both background concentrations in the Ural Region (Ural Clarke) and in the world soils (World Clarke), the total Cr concentration at urban background does not exceed the reference values, while the total Cr at anomaly sites was 2.5 times higher than Ural Clarke (Vojtkevich et al., 1977; Saet et al., 1990). Total Cr contents in podzols are known to fall into the range from 2.6 to 34 mg/kg in Canada (Frank et al., 1976) and from 3 to 200 mg/kg in the USA (Shacklette & Boerngen., 1984). The specimens with abnormal Cr concentrations (mean value was 245.2 mg/kg, maximum value was 1265.4 mg/kg) formed arbitrary spots at the study site.

4. RESULTS AND DISCUSSION

Samples gathering and chemical analysis were carried out for approximately 2 months. Models building and modelling took about two weeks. We compared two approaches to modeling the distribution of the chemical element concentrations in the surface layer of soil: geostatistical techniques (kriging), ANN, and hybrid models using MLP, GRNN and residual kriging. Quality of models prediction could be analyzed with the help of the validation (test) data set, which was not used for training networks or kriging estimates. Table 2 shows the statistical parameters used to assess the performance of the different methods (the best values are in **bold**). Figure 3 indicates the scattered diagrams of validation measurements, their estimated values at the corresponding sample sites by different methods, and linear regressions with correlation coefficients *r*.

Comparison of methods has shown the superiority of MLP and MLPRK in modeling accuracy. As Table 2 reveals, GRNNRK approach had smaller RMSE than GRNN model (1.6% improvement) and kriging (10.2% improvement). Neural networks also significantly better than kriging: GRNN and MLP have smaller RMSE (8.7%

and 9.2% respectively). At the same time, it was found that the use of a hybrid approach MLPRK gives an increase in the accuracy of prediction (based on RMSE) for about 13.0% relative to MLP and 21.0% relative to kriging, which corresponds to the results (Dai et. al., 2014). Figure 3 supports this as it shows that the GRNNRK estimates were typically less scatter compared to kriging estimates and even MLPRK predictions, and generally demonstrated a higher correlation with the sample values than kriging estimates. However, the accuracy of MLPRK was greatly better, which is supported by visual comparison in Figure 3 (b)-(d) and (c)-(e) respectively.

Thus, estimation of ANN residues by the ordinary kriging allowed smoothing out the high and low values of concentrations of the pollutant in the soil, which improves the accuracy of prediction. Consideration of regression lines (see Figure 3 (f)) shows that MLP approach provides less dispersed residues compared to GRNN, which confirms previous findings. Hence, it is possible to suggest that a large-scale variation trend of pollutant surface distribution could be quite precisely modeled by the MLP approach and residual kriging could capture its small-scale variations. Therefore, MLPRK approach demonstrated its superiority over other methods studied

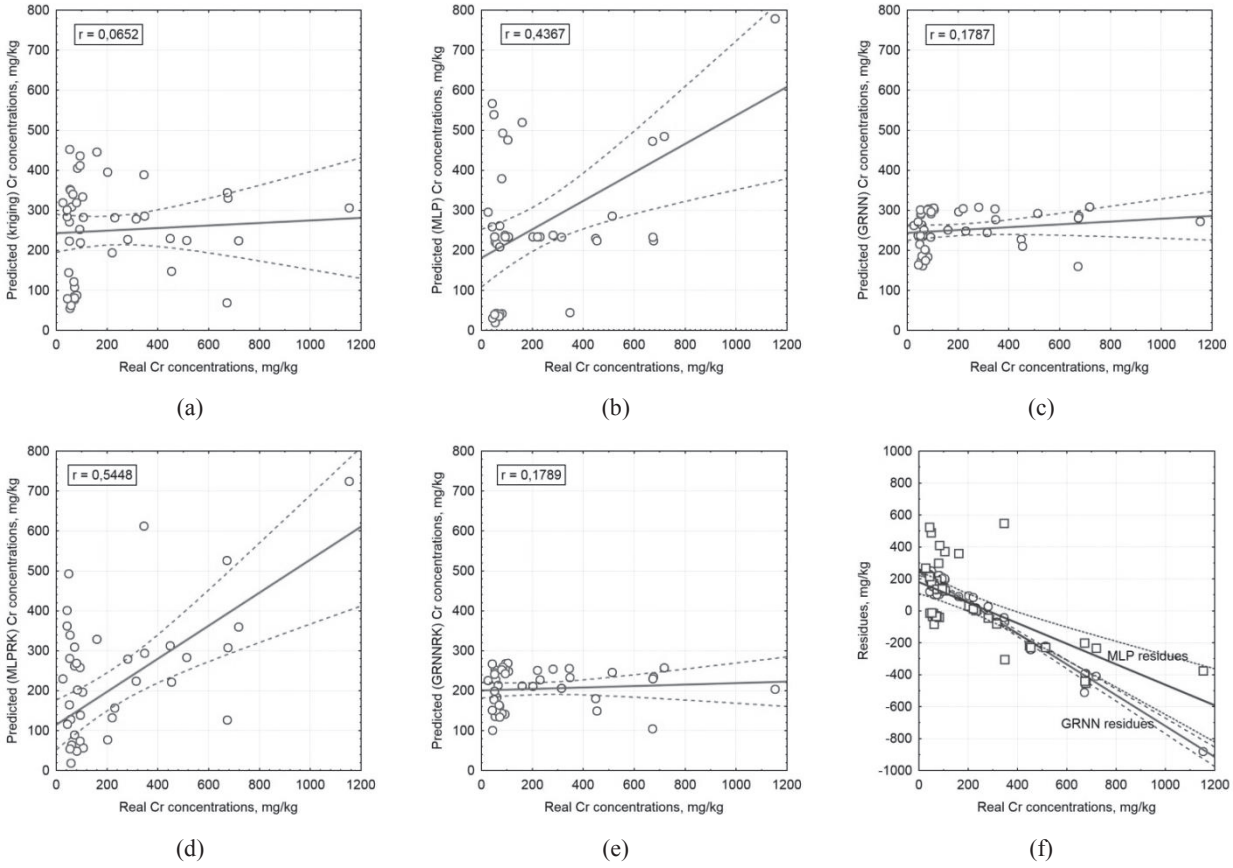

FIGURE 3. Comparison of different prediction approaches; *r* – correlation coefficient; dotted curves indicate confidence intervals: a) ordinary kriging, b) multilayer perceptron, c) generalized regression neural network, d) multilayer perceptron + residual kriging e) generalized regression neural network + residual kriging, f) GRNN and MLP models residues.

TABLE 2. Accuracy assessment indices of the pollutant (*Cr*) predicted concentrations

Method	Correlation coefficient	MAE, mg/kg	RMSE, mg/kg
Kriging	0.0652	201.88	266.36
MLP	0.4367	186.15	241.95
GRNN	0.1787	196.79	243.20
Hybrid MLP + kriging	**0.5448**	**167.52**	**210.56**
Hybrid GRNN + kriging	0.1789	176.07	239.25

5. CONCLUSION

In the present work, a modelling approach with use of the GRNNRK and MLPRK are presented. The effectiveness of these models was verified through their comparative evaluation with application of different models (kriging, MLP, GRNN, MLPRK, GRNNRK) for chromium distribution prediction in sub-Arctic Novy Urengoi, Yamalo-Nenets Autonomous Okrug, Russia. The correlation coefficient (r), the mean absolute error (MAE), and the root-mean-square error (RMSE) were used as the predictive accuracy indicators of the models for the validation (test) data set. A study on the distribution of chromium concentrations in the surface layer of soil at the urbanized terrain of the Novy Urengoi was previously conducted and described in (Sergeev et al., 2015). The aim of this study was to develop a reliable high resolution mapping model which can precisely estimate the content of an abnormally distributed soil pollutant (Cr) at a particular location.

The models were built up using the spatial coordinates as the input parameters and the chromium concentrations as the output parameters. The residues of the ANN (MLP and GRNN) models application were then analyzed for estimation the small scale variability of the data. The ordinary kriging was then performed on the residues and the outputs were combined with the ANN models to produce the MLPRK and GRNNRK models' predictions.

Thus, comparison of different approaches to the prediction of the contaminants distribution in the surface layer of soil was carried. The results showed that the MLP-based models were more accurate than the model based on the kriging and even on the GRNN. Estimation of ANN prediction residues by ordinary kriging reduced ANN prediction errors, which increased the accuracy of the model. In comparison with other methods, the most significant improvement in RMSE (21%) was observed in the MLPRK model.

The obtained results confirm vast capabilities of hybrid ANN-kriging methods that can be utilized to improve the accuracy of modeling the spatial distribution of the contaminants concentrations in the topsoil of urban areas, which characterized by high heterogeneity.

REFERENCES

1. Bishop C. (1995) Neural networks for pattern recognition. Clarendon, Oxford, 504p.
2. Dai F., Zhoua O., Lva Z., Wang X., Liu G. (2014) Spatial prediction of soil organic matter content integrating artificialneural network and ordinary kriging in Tibetan Plateau. Ecological Indicators 45, 184–194.
3. Frank, R., Ishida, K., Suda, P. (1976) Metals in agricultural soils of Ontario. Canadian Journal of Soil Science, 56, 181-196.
4. Kanevski M. F.(1999) Spatial Predictions of Soil Contamination Using General Regression Neural Networks. International Journal of Systems Research and Information Systems, vol. 8, issue 4. p.241–256.
5. Liu Z.H., Chang Y., Chen H.W. (2008) Estimation of forest volume in Huzhong forest area based on RS, GIS and ANN (in Chinese). Chinese Journal Applied Ecology 19: 1891–1896.
6. Liu F., He X. and Zhou L. (2009) Application of generalized regression neural network residual kriging for terrain surface interpolation. Proc. SPIE 7492, International Symposium on Spatial Analysis, Spatial-Temporal Data Modeling, and Data Mining, 74925F.
7. Schloeder C.A., Zimmerman N.E., Jacobs M.J. (2001) Comparison of methods for interpolating soil properties using limited data. Soil Sci. Soc. Am. J. 65, 470-479.
8. Shacklette, H. T. & Boerngen, J. G. (1984) Element concentrations in soils and other surficial materials of the conterminous United States, U.S. Geological Survey professional paper, 1270, 105.
9. Sergeev A.P., Baglaeva E.M., Shichkin A.V. (2010) Case of soil surface chromium anomaly of a northern urban territory – preliminary results, Atmospheric Pollution Research, vol. 1, 44–49.
10. Sergeev A.P., Buevich A.G., Medvedev A., Subbotina I.E., Sergeeva M. (2015) Artificial neural network and kriging interpolation for the chemical elements contents in the surface layer of soil on a background area. 15th International Multidisciplinary Scientific GeoConference SGEM 2015, Conference Proceedings, 2015, Book 3 Vol. 2. 49–56.
11. Vojtkevich, V., Miroshnikov, G.V., Boil, A.S., Prohorov, V.G. (1977) The Short Manual on Geochemistry (in Russian), Bowels, Moscow.
12. Worsham L., D. Markewitz, & N. Nibbelink (2010) Incorporating spatial dependence into estimates of soil carbon contents under different land covers. Soil Sci. Am. J. 74: 635–646.